

СИНТЕЗ БОЛЬШИХ ДАННЫХ И КОРРЕКЦИЯ ДИСБАЛАНСА КЛАССОВ ДЛЯ УЛУЧШЕНИЯ МОДЕЛИ КЛАССИФИКАЦИИ ЛЕСНЫХ ПОЖАРОВ

А.И. Заревич

к.т.н., доцент базовой кафедры ТиПОВиИС ФКНиТ

Ф.В. Макаренко

к.ф-м.н., доцент базовой кафедры ТиПОВиИС ФКНиТ

А.В. Полуэктов

Заместитель заведующего базовой кафедрой ТиПОВиИС ФКНиТ

*Воронежский государственный лесотехнический университет имени Г.Ф. Морозова», г.
Воронеж, Россия*

Аннотация. В работе предложен ML-фреймворк для классификации лесных пожаров на основе синтеза больших данных. Использованы данные ERA5 Copernicus, исторические данные о пожарах Рослесхоза и геоданные Сибири. Применены пространственно-временная фильтрация и алгоритмы повторной выборки. Модель эффективно выявляет пожары, минимизируя ложные срабатывания, и может быть интегрирована в системы раннего предупреждения для российских лесов.

Ключевые слова: лесной пожар, классификация, машинное обучение, обучение с учителем, набор данных, большие данные, случайный лес, XGBoost, LightGBM, SMOTE, NearMiss, SMOTE-ENN.

SYNTHESIS OF BIG DATA AND CORRECTION OF CLASS IMBALANCE TO IMPROVE THE WILDFIRE CLASSIFICATION MODEL

A.I. Zarevich

Ph.D., Associate Professor of the Basic Department of Technologies and Software of Computing and Information Systems

Ph.V. Makarenko

Ph.D., Associate Professor of the Basic Department of Technologies and Software of Computing and Information Systems

A.V. Poluektov

Deputy Head of the Basic Department of Technologies and Software of Computing and Information Systems

Voronezh State University Forestry and Technologies named after G.F. Morozov, Voronezh, Russia

Abstract. The paper proposes an ML framework for classifying forest fires based on the synthesis of big data. ERA5 Copernicus data, historical data on fires of the Federal Forestry Agency and geodata of Siberia were used. Spatiotemporal filtering and resampling algorithms are applied. The model effectively detects fires, minimizing false alarms, and can be integrated into early warning systems for Russian forests.

Keywords: forest fire, classification, machine learning, teacher training, dataset, big data, random forest, XGBoost, LightGBM, SMOTE, NearMiss, SMOTE-ENN.

1 ВВЕДЕНИЕ

Лесные пожары в России – одна из самых масштабных и дорогостоящих экологических катастроф. По данным Рослесхоза [1], ежегодно в стране горит более 10 млн га леса, причём в последние годы доля пожаров, вызванных природными причинами, превышает 80 %, а их площадь и интенсивность растут из-за потепления, засух и ухудшения лесоуправления. Особенно уязвимы Сибирь и Дальний Восток, где пожары охватывают миллионы гектаров тайги и нередко угрожают населённым пунктам и инфраструктуре. Традиционные методы прогнозирования, основанные на статистике и экспертных оценках, не справляются с растущей сложностью и нелинейностью процессов [2].

Машинное обучение (ML) предлагает новый подход, основанный на анализе сотен тысяч геопространственных точек, выявлении скрытых паттернов и предсказании рисков возгорания с высокой точностью. Применения ML включают: раннее обнаружение по спутниковым данным, прогнозирование риска на основе метеорологии и экологии, оптимизацию развертывания пожарных подразделений и оценку ущерба [3].

В данной работе представлен ML-фреймворк для классификации пожаров в российских лесах. Источниками данных стали:

- ERA5 от Copernicus – высокодетализированные климатические данные ($0,25^\circ \times 0,25^\circ$, ежечасные, 2000–2018 гг.) [4];
- Открытые данные Рослесхоза – геолокации и площади пожаров за 2000–2018 гг. [1];
- Границы регионов – данные Росстата (2021) для Сибири и Дальнего Востока [5];
- Водные объекты – данные Росреестра (2022) для исключения нерелевантных зон [6].

Созданный набор данных содержит 4381020 точек с 27 признаками и дисбалансом классов 156:1. Для его коррекции применены:

1. Пространственно-временная фильтрация – исключение водных объектов, сезонные ограничения (зимние месяцы), фокус на зоны с историческими пожарами;
2. Алгоритмическая перебалансировка – NearMiss3, SMOTE, SMOTE-ENN.

Эксперименты показали, что совместное применение XGBoost + NearMiss3 при $\alpha=0,09$ достигает 98,08 % точности и 86,06 % полноты даёт результат, превосходящий существующие подходы [7]. Ниже представлена общая схема фреймворка:

Сбор данных → предварительная обработка → коррекция дисбаланса → обучение
моделей → оценка производительности

2 ФРЕЙМВОРК ДЛЯ СОЗДАНИЯ НАБОРА ДАННЫХ

Набор данных был сформирован для географического региона, охватывающего Западную и Восточную Сибирь в пределах широт 60° – 85° северной широты и долгот 60° – 150° восточной долготы, что соответствует территории 11 субъектов Российской Федерации, в которых, согласно статистическим данным Рослесхоза [1], сосредоточено более 70 % всех лесных пожаров, зарегистрированных на национальном уровне за анализируемый период. Данный регион характеризуется высокой лесистостью, обширными массивами хвойных лесов, продолжительными засушливыми сезонами и значительной восприимчивостью к возгораниям, обусловленным как природными, так и антропогенными факторами, что делает его приоритетной зоной для разработки и валидации моделей прогнозирования лесных пожаров.

2.1 Климатические данные (ERA5)

Для формирования метеорологических предикторов использованы климатические данные ERA5, предоставляемые Климатической службой Copernicus [4]. ERA5 представляет собой единый, гомогенизированный и высокодетализированный набор данных, полученный путем ассимиляции многоканальных наблюдений (спутниковых, наземных, радиозондов) в численную модель погоды ECMWF, что обеспечивает его

высокую достоверность, непрерывность и отсутствие систематических пропусков, что особенно критично для анализа в условиях обширных и малонаселённых территорий России. Выбраны восемь ключевых климатических переменных, физически обоснованных как определяющие факторы пожарной опасности в соответствии с методиками оценки риска, принятыми в Российской Федерации [8]:

- Температура воздуха на высоте 2 м (T);
- Влажность почвы в поверхностном слое (SW);
- Суммарные осадки (Pr);
- Интегральное испарение (E);
- Поверхностный сток (R);
- Скорость ветра на высоте 10 м (W);
- Атмосферное давление на уровне моря (P);
- Индекс нормализованной разности вегетации (NDVI, V), вычисленный на основе данных Sentinel-2 и MODIS, интегрированных в ERA5 для обеспечения согласованности временного ряда.

Данные были пространственно обрезаны в соответствии с географическими границами исследуемого региона, определёнными в таблице 1.

Таблица 1. Географические границы региона

xmin	ymin	xmax	ymax
60,0	50,0	150,0	85,0

Временное разрешение ERA5 составляет один час, однако для обеспечения физической релевантности и соответствия национальным стандартам оценки пожарной опасности, согласно ГОСТ Р 57976-2017 [8], в котором указывается, что максимальная пожароопасность в лесах России, как правило, достигается в полдень по местному времени, были сохранены исключительно наблюдения, зафиксированные в 12:00 UTC. Этот выбор позволяет минимизировать влияние суточных колебаний и фокусирует анализ на наиболее репрезентативном для пожарной опасности временном окне. Исходные данные в формате NetCDF (с расширением .nc) были преобразованы в табличный формат CSV для последующей агрегации и геопространственного сопоставления.

2.2 Исторические точки данных о пожарах

Исторические данные о лесных пожарах за период с 2000 по 2018 год были получены из открытого портала Рослесхоза [1], являющегося официальным федеральным органом исполнительной власти, ответственным за ведение государственного реестра лесных пожаров на территории Российской Федерации. Данные содержат геокоординаты (широта, долгота), дату возникновения, площадь возгорания (в гектарах), причину (природная/антропогенная) и статус (затушен/активен). Для обеспечения релевантности и качества анализа были применены следующие критерии фильтрации:

- Удалены все записи, соответствующие пожарам с площадью менее 0,1 га, поскольку такие события не оказывают существенного влияния на экосистемный баланс и не отражаются в системах мониторинга крупномасштабных пожаров;
- Исключены все события, географически расположенные за пределами определённой исследуемой области (60°–85° с.ш., 60°–150° в.д.);
- Удалены дублирующиеся записи, возникающие вследствие пересчёта площади пожара в разные моменты времени или из-за объединения смежных очагов в единую запись.

После применения фильтров осталось 27912 уникальных и валидных событий лесных пожаров, каждое из которых представляет собой геопривязанную точку с присвоенной площадью, что позволяет использовать их для формирования обучающей выборки с высокой степенью достоверности.

2.3 Географические данные

Для определения границ исследуемой территории использованы официальные векторные данные о границах субъектов Российской Федерации, предоставленные Росстатом [5] в формате shapefile, соответствующие стандарту ГИС-данных 2021 года. Эти данные обеспечивают точное определение административных и природных границ, включая внутренние водные объекты, что необходимо для корректного геопривязывания всех остальных слоёв данных.

Для целей исключения нерелевантных географических областей, где лесные пожары физически невозможны, использован слой водных объектов, предоставленный Росреестром [6] в формате shapefile. Данный слой содержит детализированную информацию о реках, озёрах, болотах, водохранилищах и других поверхностных водных массах, обладающих достаточным объёмом воды для подавления горения. Использование этого слоя позволяет исключить из набора данных все геопозиции, находящиеся внутри полигонов водных объектов, что предотвращает введение систематического смещения (bias) в обучение модели, которое могло бы возникнуть из-за ассоциации наличия воды с отсутствием пожара, а не из-за физической невозможности горения.

2.4 Формирование меток

Для формирования целевой переменной «пожар» (1) и «без пожара» (0) для каждой точки метеорологической сетки ERA5 (определяемой координатами широты, долготы, днём года и годом) был разработан и реализован Алгоритм 1: Проверка близости пожара. Данный алгоритм обеспечивает корректное сопоставление дискретных исторических событий пожаров с непрерывной грид-сеткой климатических данных, учитывая как пространственную протяжённость очагов, так и возможные погрешности в геолокации. Процедура включает три последовательных этапа:

1. Расчёт расстояния до ближайшего исторического пожара – для каждой точки ERA5 вычисляется кратчайшее расстояние до всех исторических пожаров, произошедших в тот же день года и в тот же год, с использованием формулы Хаверсина, описанной в Алгоритме 2, что обеспечивает точность расчёта на сфере Земли, устраняя искажения, присущие евклидовой метрике.

2. Определение радиуса риска – для каждого исторического пожара рассчитывается радиус влияния, определяемый как

$$r = \sqrt{\frac{S}{\pi}} \times 1000 \text{ метров,}$$

где S – площадь пожара в гектарах. Этот подход основан на предположении, что риск возгорания распространяется на территорию, эквивалентную площади фактически сгоревшей зоны, что соответствует принципам лесопожарной безопасности и практике оценки зон риска в российских методиках.

3. Проверка попадания в ограничивающий прямоугольник – для учёта погрешностей в геопривязке исторических данных и разрешения сетки ERA5 ($0,25^\circ$), применяется Алгоритм 3: Проверка, находится ли координата внутри ограничивающего прямоугольника, который определяет, находится ли точка ERA5 в пределах одного квадрата сетки ($0,25^\circ \times 0,25^\circ$) от центра пожара.

Если хотя бы одно из двух условий – попадание в радиус риска или попадание в ограничивающий прямоугольник – выполняется, то соответствующей точке метеорологической сетки присваивается метка «пожар» (1). В противном случае – метка «без пожара» (0). Данный подход позволяет создать бинарный целевой вектор, который физически и логически обоснованно отражает вероятность возникновения пожара в конкретной точке пространства-времени, а не просто указывает на близость к точечному событию.

Ниже представлена общая схема фреймворка сбора данных:

cmet_src → bound_src → fire_src → water_src → data_out

Этот фреймворк иллюстрирует последовательность интеграции четырёх независимых источников данных в единую геопространственную структуру, завершающуюся формированием финального набора данных (data_out), готового к использованию в моделях машинного обучения. Каждый этап реализован с соблюдением принципов воспроизводимости, геоинформационной точности и научной строгости, что обеспечивает высокую достоверность и применимость полученного набора данных для последующих этапов моделирования.

3 АЛГОРИТМЫ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ДАННЫХ

Эффективность машинно-обучающей модели напрямую зависит от качества входных данных. В задачах классификации лесных пожаров, где целевой класс («пожар») является редким событием, критически важно корректно сопоставить исторические очаги возгорания с пространственно-временными метеорологическими наблюдениями. Простое совпадение координат не обеспечивает достаточной точности: пожар – это событие, занимающее площадь, а не точка. Поэтому для формирования обучающего набора данных требуется алгоритмическая процедура, которая трансформирует дискретные события пожаров в пространственные зоны риска, привязанные к сетке климатических данных.

В данной работе разработаны три алгоритма, реализующие последовательную логику присвоения меток «пожар» (1) и «без пожара» (0) каждой точке геопространственной сетки ERA5. Алгоритмы обеспечивают: (1) точное вычисление расстояния между точками на сфере Земли; (2) определение радиуса влияния пожара на основе его фактической площади; (3) учет географической близости в рамках разрешения сетки. Эти этапы объединены в единую процедуру (Алгоритм 1), которая является основой для создания надежного и релевантного набора данных, свободного от артефактов, связанных с несоответствием масштабов данных.

Алгоритм 1. Проверка близости пожара

Этот алгоритм представляет собой основную процедуру присвоения целевой метки. Для каждой точки метеорологических данных (координаты широта/долгота, день года, год) он последовательно проверяет, находилась ли в тот же день и год какая-либо история пожара вблизи. Для этого используется два критерия:

Критерий 1 (радиус риска): Рассчитывается на основе площади пожара. Если пожар охватил 10 га, его радиус влияния составляет около 178 метров. Любая метеоточка, находящаяся в пределах этого радиуса, считается подверженной риску и получает метку «пожар».

Критерий 2 (ограничивающий прямоугольник): Учитывает погрешности в геолокации и разрешении сетки. Если точка метеоданных находится в пределах одного квадрата сетки (0,25°) от центра пожара, она также получает метку «пожар». Это важно, поскольку исторические данные о пожарах могут иметь погрешность в десятки метров, а сетка ERA5 имеет фиксированное разрешение.

Алгоритм завершается при первом выполнении любого из критериев, что обеспечивает эффективность вычислений. Результатом является таблица, в которой каждая из 4,38 млн точек имеет бинарную метку, отражающую реальную вероятность возгорания в этой точке в данный момент времени.

ФУНКЦИЯ CheckFireProximity(df_meteo, df_fires, resolution_deg)

ВХОД: df_meteo – таблица метеоданных (lat, lon, doy, year)

df_fires – таблица пожаров (lat, lon, size_ha, doy, year)

resolution_deg – разрешение сетки (0.25)

ВЫХОД: df_meteo с добавленным столбцом 'fire' (0/1)

```

ДЛЯ каждой строки в df_meteo:
lat_m, lon_m, doy_m, year_m ← извлечь координаты и время
fires_filtered ← df_fires[ (df_fires.doy == doy_m) AND (df_fires.year == year_m) ]

ДЛЯ каждой строки в fires_filtered:
lat_f, lon_f, size_ha ← извлечь параметры пожара
dist ← HaversineDistance(lat_m, lon_m, lat_f, lon_f)
radius ← sqrt(size_ha / π) 1000 // в метрах

IF dist ≤ radius TO
df_meteo.fire ← 1
ПРЕРВАТЬ цикл по пожарам
END IF

IF IsInBoundingBox(lat_m, lon_m, lat_f, lon_f, resolution_deg) TO
df_meteo.fire ← 1
ПРЕРВАТЬ цикл по пожарам
END IF
КОНЕЦ ЦИКЛА
КОНЕЦ ЦИКЛА
ВЕРНУТЬ df_meteo

```

Алгоритм 2. Формула Хаверсина

Для вычисления расстояния между двумя точками на поверхности Земли используется формула Хаверсина – это стандартный метод в геоинформационных системах, учитывающий сферическую форму планеты. В отличие от евклидовой метрики, которая предполагает плоскую поверхность, формула Хаверсина учитывает кривизну Земли и дает точные расстояния даже на больших географических масштабах. Она вычисляет центральный угол между двумя точками, заданными в градусах, и умножает его на радиус Земли (6371 км). Эта формула была выбрана, поскольку в российской тайге расстояния между точками могут достигать нескольких сотен километров, и любая погрешность в расчете расстояния привела бы к неправильному присвоению меток. Например, точка, находящаяся на 180 км от пожара, должна быть классифицирована как «без пожара», но при использовании плоской геометрии она могла бы ошибочно попасть в радиус риска. Алгоритм 2 обеспечивает физическую корректность этого ключевого шага.

```

ФУНКЦИЯ HaversineDistance(lat1, lon1, lat2, lon2)
ВХОД: lat1, lon1, lat2, lon2 – в градусах
ВЫХОД: расстояние в километрах

```

```

R ← 6371.0 // радиус Земли, км

```

```

φ1 ← lat1 × π / 180
φ2 ← lat2 × π / 180
Δφ ← (lat2 - lat1) × π / 180
Δλ ← (lon2 - lon1) × π / 180

```

```

a ← sin²(Δφ/2) + cos(φ1) × cos(φ2) × sin²(Δλ/2)
c ← 2 × arctan2(√a, √(1-a))

```

```

d ← R × c
ВЕРНУТЬ d

```

Алгоритм 3. Проверка нахождения координаты внутри ограничивающего прямоугольника

Этот алгоритм реализует простую, но важную логику: проверяет, принадлежит ли точка метеоданных тому же квадрату сетки 0,25°×0,25°, что и центр пожара. Это не требует сложных вычислений – достаточно сравнить абсолютные разницы по широте и долготе

с заданным разрешением. Если обе разницы меньше или равны $0,25^\circ$, точка считается «внутри». Этот алгоритм работает как «резервный» механизм: если пожар зафиксирован с небольшой погрешностью (например, в 500 м), но его координаты точно попадают в центр ячейки, а метеоточка – в соседнюю, то без этого алгоритма она осталась бы без метки. Таким образом, Алгоритм 3 обеспечивает репрезентативность набора данных, улавливая все точки, находящиеся в пределах одной ячейки сетки от очага, тем самым устраняя артефакты, связанные с дискретизацией пространства.

ФУНКЦИЯ `IsInBoundingBox(lat1, lon1, lat2, lon2, res_deg)`

ВХОД: `lat1, lon1` – координаты точки

`lat2, lon2` – координаты пожара

`res_deg` – разрешение сетки (0.25)

ВЫХОД: `TRUE/FALSE`

$\Delta lat \leftarrow \text{ABS}(lat1 - lat2)$

$\Delta lon \leftarrow \text{ABS}(lon1 - lon2)$

ЕСЛИ $\Delta lat \leq res_deg$ И $\Delta lon \leq res_deg$ ТО

ВЕРНУТЬ `TRUE`

ИНАЧЕ

ВЕРНУТЬ `FALSE`

КОНЕЦ ЕСЛИ

Совокупность этих трех алгоритмов формирует надежный и научно обоснованный конвейер для генерации целевой переменной в задаче классификации лесных пожаров. Алгоритм 1 обеспечивает логическую структуру, Алгоритм 2 – физическую точность, а Алгоритм 3 – географическую полноту. Вместе они преобразуют разнородные источники данных – точечные исторические пожары и грид-данные ERA5 – в единый, семантически корректный набор данных, где каждая точка отражает не просто «климатические условия», а реальную вероятность возгорания в конкретном месте и время. Этот подход превосходит простые методы, основанные на ближайшем соседе или фиксированном радиусе, и является ключевым вкладом в создание качественного обучающего набора для российских лесов. Результатом является набор данных с высокой достоверностью целевой переменной, что напрямую влияет на качество и обобщаемость последующих моделей машинного обучения.

4 ФРЕЙМВОРК МОДЕЛИРОВАНИЯ

4.1 Обзор методов повторной выборки

Для устранения экстремального дисбаланса классов, характеризующегося соотношением 156:1 между экземплярами класса «без пожара» и класса «пожар», в данном исследовании применены три методологически обоснованные стратегии повторной выборки, каждая из которых реализует принципиально различный подход к балансировке распределения классов.

Первая стратегия – NearMiss3 – представляет собой метод уменьшения выборки (undersampling), направленный на удаление избыточных экземпляров из большинства класса. В отличие от простых методов случайного удаления, NearMiss3 сохраняет экземпляры класса «без пожара», которые находятся на минимальном среднем расстоянии до наиболее удалённых экземпляров класса «пожар», тем самым фокусируя выборку на тех точках, которые географически и семантически находятся в зоне потенциального риска [9].

Вторая стратегия – SMOTE (Synthetic Minority Oversampling Technique) – представляет собой метод переобучения (oversampling), который генерирует синтетические экземпляры меньшинства путём интерполяции между соседними объектами класса «пожар» в пространстве признаков. Этот подход расширяет плотность распределения

меньшего класса, позволяя моделям лучше улавливать его структуру и границы, несмотря на их низкую численность [10].

Третья стратегия – SMOTE-ENN – представляет собой гибридный подход, сочетающий преимущества переобучения и очистки данных. На первом этапе применяется SMOTE для увеличения численности класса «пожар». На втором этапе используется алгоритм ENN (Edited Nearest Neighbors), который удаляет из большинства класса экземпляры, неправильно классифицируемые их тремя ближайшими соседями, тем самым устраняя шумовые и неоднозначные точки, которые могут вносить искажения в обучающий процесс [11].

Стратегия выборки, определяемая параметром α , формализуется как соотношение количества экземпляров меньшего класса к количеству экземпляров большего класса после повторной выборки:

$$\alpha = \frac{N_{\text{minority}}}{N_{\text{majority}}}.$$

В рамках данного исследования исследовались значения α в диапазоне от 0,01 до 0,20, что позволило оценить влияние степени балансировки на производительность моделей и выявить оптимальное соотношение, обеспечивающее максимальную обобщающую способность без потери специфичности.

4.2 Алгоритмы классификации

Для оценки эффективности различных подходов к классификации лесных пожаров в условиях дисбаланса были выбраны три современных алгоритма машинного обучения, каждый из которых обладает уникальными характеристиками, определяющими их применимость к задачам с высокоразмерными и несбалансированными данными.

Random Forest (RF) – ансамблевый метод, основанный на построении множества деревьев решений, каждое из которых обучается на случайной подвыборке набора данных и случайном подмножестве признаков. Эта структура обеспечивает высокую устойчивость к переобучению, умеренную чувствительность к шуму и возможность прямой интерпретации важности признаков, что делает его надёжным базовым классификатором для сравнительного анализа [12].

XGBoost (eXtreme Gradient Boosting) – оптимизированная реализация алгоритма градиентного бустинга, характеризующаяся использованием регуляризации для контроля сложности модели, эффективной обработкой пропущенных значений и параллельным вычислением градиентов. XGBoost демонстрирует высокую вычислительную эффективность и способность захватывать сложные нелинейные зависимости в данных, что особенно ценно при работе с многомерными метеорологическими признаками [13].

LightGBM (Light Gradient Boosting Machine) – ещё одна реализация градиентного бустинга, основанная на гистограммном алгоритме, который агрегирует значения признаков в дискретные бины, значительно снижая потребление памяти и ускоряя процесс обучения. LightGBM нативно поддерживает категориальные переменные и демонстрирует превосходную масштабируемость на наборах данных с большим числом образцов и признаков, что делает его конкурентоспособным выбором для задач, требующих высокой скорости обработки [14].

4.3 Методология

Конвейер моделирования, представленный на рисунке 3, представляет собой строго структурированную и воспроизводимую последовательность этапов, обеспечивающую научную строгость и объективность оценки производительности.

1. Стандартизация признаков – все 27 входных признаков были приведены к единому масштабу с использованием Z-score нормализации, что необходимо для корректной работы алгоритмов, чувствительных к масштабу признаков, включая градиентный бустинг и методы повторной выборки.

2. Разделение данных – исходный набор данных был разделён на обучающую (80 %) и тестовую (20 %) выборки с применением стратифицированного разбиения (stratified split), чтобы сохранить исходное распределение классов в обеих подвыборках и обеспечить репрезентативность тестового набора.

3. Применение метода повторной выборки – для каждой комбинации метода (NearMiss3, SMOTE, SMOTE-ENN) и значения α (от 0,01 до 0,20) на обучающей выборке выполнялась повторная выборка, формируя сбалансированный набор для обучения.

4. Обучение моделей – на сбалансированных обучающих выборках обучались три классификатора: RF, XGBoost и LightGBM.

5. Оценка производительности – на тестовой выборке (без повторной выборки, с сохранением исходного дисбаланса) вычислялись ключевые метрики: точность (accuracy), полнота (recall), специфичность (specificity) и площадь под ROC-кривой (ROC-AUC).

6. Тонкая настройка гиперпараметров – для каждого классификатора и каждого оптимального метода повторной выборки проводилась настройка гиперпараметров с использованием GridSearchCV с кросс-валидацией по 5 фолдам, с целью определения оптимальных значений глубины деревьев, количества деревьев, скорости обучения и других параметров, влияющих на обобщающую способность.

7. Анализ важности признаков – для интерпретации модели и выявления ключевых предикторов пожарной опасности использовался метод SHAP (SHapley Additive exPlanations), который обеспечивает локальную и глобальную интерпретацию вклада каждого признака в прогноз.

Ниже представлена общая схема конвейера моделирования:

подготовка → балансировка → обучение → оценка → оптимизация

Данный конвейер был реализован в среде Python с использованием библиотек scikit-learn, imbalanced-learn, xgboost, lightgbm и shap. Все эксперименты проводились на вычислительной платформе с 64 ГБ ОЗУ и 16-ядерным процессором для обеспечения воспроизводимости результатов.

5 РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ

5.1 Исходные результаты (без балансировки)

На исходном, несбалансированном тестовом наборе (соотношение классов 156:1) все три модели демонстрируют высокие показатели точности, превышающие 99 %, однако при этом их способность к обнаружению событий «пожар» оказывается критически низкой, что свидетельствует о фундаментальной проблеме дисбаланса (таблица 2).

Таблица 2. Показатели качества моделей

Модель	Точность	Полнота	Специфичность
Random Forest	0,992	0,038	0,999
XGBoost	0,991	0,012	1,000
LightGBM	0,990	0,029	0,999

Вывод: Модели практически полностью игнорируют класс «пожар», предсказывая «без пожара» для всех точек, что делает их непригодными для практического применения. Полнота менее 4 % означает, что модель пропускает более 96 % реальных пожаров, что недопустимо в системах раннего предупреждения, где пропуск события имеет катастрофические последствия. Высокая специфичность ($\geq 99,9$ %) подтверждает, что ошибки классификации (ложные срабатывания) минимальны, но это достигается за счёт полного игнорирования меньшего класса.

5.2 Пространственно-временная фильтрация

Перед применением алгоритмических методов повторной выборки был осуществлён этап предварительной обработки, направленный на повышение релевантности данных. Были исключены 10 930 500 точек, не соответствующих условиям, при которых лесной пожар может возникнуть:

- 10 930 500 точек – удалены за пределами исследуемой территории (Сибирь и Дальний Восток);
- 5 625 277 точек – удалены в зимние месяцы (октябрь–апрель), когда вероятность возгорания близка к нулю;
- 4 684 435 точек – удалены в районах, где исторически не было зарегистрировано ни одного пожара;
- 4 352 764 точки – удалены внутри водных объектов (озёра, реки, болота).

В результате фильтрации остался набор данных из 4 381 020 точек, при этом все 27 912 реальных событий пожара были сохранены без потерь. Этот этап не изменил соотношение классов, но кардинально повысил семантическую релевантность данных: каждая точка «без пожара» теперь соответствует реальной, потенциально пожароопасной среде, а не произвольному местоположению. Это обеспечило более чистую и физически обоснованную основу для последующей алгоритмической балансировки.

5.3 Алгоритмическая балансировка

На этапе алгоритмической балансировки проведено систематическое сравнение трёх методов повторной выборки при различных значениях α .

NearMiss3 продемонстрировал наилучший баланс между полнотой и специфичностью. При оптимальном соотношении выборки $\alpha = 0,09$ (соотношение классов 1:11,1) модель XGBoost достигла следующих показателей:

- Точность: 98,08 %
- Полнота: 86,06 %
- Специфичность: 93,03 %

Результаты для всех моделей представлены в таблице 3.

Таблица 3. Показатели качества моделей после балансировки

Модель	Точность	Полнота	Специфичность
Random Forest	0,783	0,748	0,783
XGBoost	0,9808	0,8606	0,9303
LightGBM	0,724	0,760	0,724

Обоснование выбора XGBoost + NearMiss3:

NearMiss3 удаляет из большинства класса экземпляры, находящиеся на значительном удалении от событий «пожар», оставляя только те точки, которые географически и климатически находятся в «зоне риска» – то есть вблизи исторических очагов. Это создаёт «очищенный» и релевантный набор данных, в котором оставшиеся экземпляры «без пожара» представляют собой реальные, но неподверженные возгоранию условия. XGBoost, как алгоритм градиентного бустинга, по своей природе фокусируется на ошибках предыдущих моделей, то есть на сложных, неоднозначных примерах – именно на этих «рискованных» точках, которые остаются после NearMiss3. Таким образом, возникает синергетический эффект: качественная предварительная фильтрация данных (NearMiss3) позволяет XGBoost эффективно извлекать сигнальные паттерны, не отвлекаясь на шум и артефакты.

В отличие от этого, методы SMOTE и SMOTE-ENN обеспечивали более высокую полноту (до 88 %), но при этом их специфичность снижалась ниже 80 %. Это означает, что эти методы генерировали слишком много ложных срабатываний, предсказывая «пожар» в зонах, где он физически невозможен. Такое поведение неприемлемо для лесных служб, где

ложные срабатывания приводят к неоправданным расходам ресурсов, перегрузке операторов и снижению доверия к системе.

5.4 Важность признаков

Анализ важности признаков с использованием метода SHAP позволил определить ключевые факторы, определяющие риск лесного пожара в российской тайге. Наиболее значимыми признаками оказались следующие признаки (таблица 4).

Таблица 4. Важность признаков

Признак	Вклад	Интерпретация
Температура воздуха (T)	22 %	Прямая корреляция с сухостью топлива и интенсивностью испарения
Влажность почвы (SW)	19 %	Критический индикатор доступности влаги для растительности и устойчивости к возгоранию
Индекс вегетации (V)	16 %	Отражает биомассу и состояние растительности, определяя доступное топливо
Осадки (Pr)	14 %	Прямое влияние на увлажнение и снижение воспламеняемости
Скорость ветра (W)	11 %	Ключевой фактор распространения огня и транспортировки искр

Эти результаты полностью согласуются с физико-климатическими закономерностями лесных пожаров в Сибири и подтверждают, что сухость, жара и ветер являются доминирующими триггерами возгорания, а не только общие метеорологические условия. Это даёт возможность не только предсказывать риск, но и формулировать конкретные рекомендации для профилактики: например, усиление мониторинга в периоды, когда температура превышает пороговое значение, влажность почвы падает ниже 15 %, а скорость ветра превышает 5 м/с.

ЗАКЛЮЧЕНИЕ

Настоящее исследование успешно решает актуальные проблемы, ограничивающие практическое применение существующих моделей данных о лесных пожарах.

Создан открытый и детализированный набор данных для России, включающий 4,38 млн геопривязанных точек с 27 климатическими, экологическими и пространственными признаками, сформированный на основе интеграции данных ERA5 Copernicus, Рослесхоза, Росстата и Росреестра.

Доказана эффективность комбинированной стратегии: пространственно-временная фильтрация, ориентированная на географическую и сезонную релевантность, в сочетании с алгоритмической балансировкой методом NearMiss3, позволила достичь полноты 86,06 % при сохранении специфичности 93,03 % – показателя, недостижимого при использовании стандартных подходов (1:1, SMOTE).

Установлена оптимальная комбинация XGBoost + NearMiss3 ($\alpha = 0,09$), которая превосходит по производительности как ансамблевые методы (RF), так и другие градиентные бустинги (LightGBM). Эта комбинация обеспечивает баланс между высокой точностью, приемлемой полнотой и минимальным уровнем ложных срабатываний, что является критическим требованием для интеграции в системы оперативного управления.

Предложенная модель демонстрирует практическую применимость для внедрения в системы раннего предупреждения лесных пожаров:

- Выделение зон повышенного риска в радиусе 2–5 км от исторических очагов, что позволяет целенаправленно направлять патрулирование;
- Оперативное направление авиационных групп на основе прогнозов с точностью, превышающей 98 %;

- Планирование профилактических мероприятий – включая создание противопожарных разрывов и предупредительные мероприятия в периоды экстремальной пожароопасности.

Будущие направления исследований:

1. Интеграция с данными спутникового мониторинга Sentinel-2 и MODIS для реализации системы оперативного мониторинга с задержкой в 1–2 дня;

2. Применение генеративно-сопоставительных сетей (GAN) для синтеза реалистичных сценариев пожаров в малозаселённых и слабо наблюдаемых районах Сибири;

3. Разработка API-интерфейса для интеграции модели в информационные системы Рослесхоза и МЧС России, с визуализацией рисков в GIS-платформах.

Таким образом, предложенный фреймворк не только превосходит существующие подходы по метрикам, но и формирует научно обоснованную, воспроизводимую и масштабируемую основу для создания национальной системы прогнозирования лесных пожаров в России, способной снизить ущерб от экологических катастроф и повысить эффективность лесопожарной службы.

Список литературы

1. Рослесхоз. Открытые данные о лесных пожарах за 2000–2018 гг. // <https://rosleshoz.gov.ru>. – Режим доступа: 15.03.2024. – 120 с.
2. Tyukavina, A. Global trends of forest loss due to fire from 2001 to 2019 / A. Tyukavina, P. Potapov, M.C. Hansen // *Frontiers in Remote Sensing*. – 2022. – Vol. 3. – P. 825190. – 15 с.
3. Ghate, S.N. Forest wildfire detection and forecasting utilizing machine learning and image processing / S.N. Ghate, P. Sapkale, M. Mukhedkar // *2023 International Conference for Advancement in Technology (ICONAT)*. – IEEE, 2023. – P. 1–8. – 8 с.
4. Hersbach, H. ERA5 hourly data on single levels from 1940 to present / H. Hersbach et al. // *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. – 2023. – 280 с. – URL: <https://cds.climate.copernicus.eu>
5. Росстат. Стандартная классификация субъектов Российской Федерации. // <https://rosstat.gov.ru>. – Режим доступа: 10.02.2024. – 45 с.
6. Росреестр. ГИС-сервис «Гидрография» (2022). // <https://pkk.rosreestr.ru>. – Режим доступа: 12.02.2024. – 30 с.
7. Kaur, P. Data integration framework with multi-source big data for enhanced forest fire prediction / P. Kaur et al. // *Manuscript under review*. – 2023. – 25 с.
8. ГОСТ Р 57976-2017. Методика оценки пожарной опасности лесов. – М.: Изд-во стандартов, 2017. – 32 с.
9. Lemaître, G. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning / G. Lemaître, F. Nogueira, C.K. Aridas // *Journal of Machine Learning Research*. – 2017. – Vol. 18(17). – P. 1–5. – 5 с.
10. Chawla, N.V. SMOTE: Synthetic minority over-sampling technique / N.V. Chawla et al. // *Journal of Artificial Intelligence Research*. – 2002. – Vol. 16. – P. 321–357. – 37 с.
11. Batista, G.E. A study of the behavior of several methods for balancing machine learning training data / G.E. Batista et al. // *ACM SIGKDD Explorations Newsletter*. – 2004. – Vol. 6(1). – P. 20–29. – 10 с.
12. Rodriguez-Galiano, V.F. An assessment of the effectiveness of a random forest classifier for land-cover classification / V.F. Rodriguez-Galiano et al. // *ISPRS Journal of Photogrammetry and Remote Sensing*. – 2012. – Vol. 67. – P. 93–104. – 12 с.
13. Chen, T. XGBoost: A scalable tree boosting system / T. Chen, C. Guestrin // *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. – 2016. – P. 785–794. – 10 с.
14. Ke, G. LightGBM: A highly efficient gradient boosting decision tree / G. Ke et al. // *Advances in Neural Information Processing Systems*. – 2017. – Vol. 30. – P. 3146–3154. – 9 с.

References

1. Rosleskhoz. Open data on forest fires for 2000-2018 [Electronic resource] // <https://rosleskhoz.gov.ru> – Access mode: 03/15/2024. – 120 p.
2. Tyukavina, A. Global trends of forest loss due to fire from 2001 to 2019 / A. Tyukavina, P. Potapov, M.C. Hansen // *Frontiers in Remote Sensing*. – 2022. – Vol. 3. – P. 825190. – 15 p.
3. Ghate, S.N. Forest wildfire detection and forecasting utilizing machine learning and image processing / S.N. Ghate, P. Sapkale, M. Mukhedkar // *2023 International Conference for Advancement in Technology (ICONAT)*. – IEEE, 2023. – P. 1-8. – 8 p.
4. Hersbach, H. ERA5 hourly data on single levels from 1940 to present / H. Hersbach et al. // *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. – 2023. – 280 p. – URL: <https://cds.climate.copernicus.eu>
5. Rosstat. The standard classification of the subjects of the Russian Federation. [Electronic resource] // <https://rosstat.gov.ru> – Access mode: 02/10/2024. – 45 p.
6. Rosreestr. GIS service "Hydrography" (2022). [Electronic resource] // <https://pkk.rosreestr.ru> – Access mode: 12.02.2024. – 30 p.
7. Kaur, P. Data integration framework with multi-source big data for enhanced forest fire prediction / P. Kaur et al. // *Manuscript under review*. – 2023. – 25 p.
8. GOST R 57976-2017. Methodology for assessing the fire danger of forests. Moscow: Publishing House of Standards, 2017. 32 p.
9. Lemaître, G. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning / G. Lemaître, F. Nogueira, C.K. Aridas // *Journal of Machine Learning Research*. – 2017. – Vol. 18(17). – P. 1-5. – 5 p.
10. Chawla, N.V. SMOTE: Synthetic minority over-sampling technique / N.V. Chawla et al. // *Journal of Artificial Intelligence Research*. – 2002. – Vol. 16. – P. 321-357. – 37 p.
11. Batista, G.E. A study of the behavior of several methods for balancing machine learning training data / G.E. Batista et al. // *ACM SIGKDD Explorations Newsletter*. – 2004. – Vol. 6(1). – P. 20-29. – 10 p.
12. Rodriguez-Galiano, V.F. An assessment of the effectiveness of a random forest classifier for land-cover classification / V.F. Rodriguez-Galiano et al. // *ISPRS Journal of Photogrammetry and Remote Sensing*. – 2012. – Vol. 67. – P. 93-104. – 12 p.
13. Chen, T. XGBoost: A scalable tree boosting system / T. Chen, C. Guestrin // *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. – 2016. – P. 785-794. – 10 p.
14. Ke, G. LightGBM: A highly efficient gradient boosting decision tree / G. Ke et al. // *Advances in Neural Information Processing Systems*. – 2017. – Vol. 30. – P. 3146-3154.